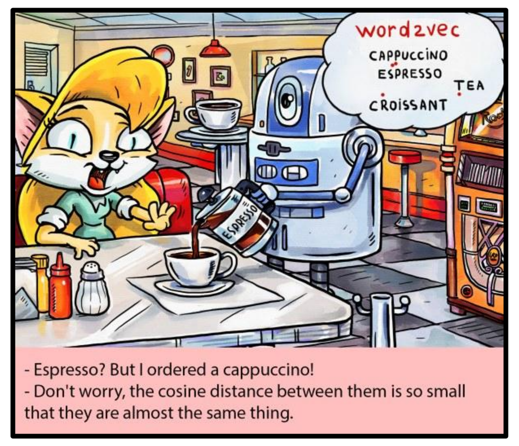


Rep4NLP: 2nd Workshop on Representation Learning for NLP

Regularized Topic Models for Sparse Interpretable Word Embeddings

Anna Potapenko[1] [apotapenko.com], Artem Popov[2], Konstantin Vorontsov[1,2]
[1] National Research University Higher School of Economics; [2] Lomonosov Moscow State University, Moscow, Russia

Topic Models

- Text mining tool to reveal hidden topics
- Utilize word-document co-occurrence data
- PLSA model:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td},$$

  - Parametrized by two matrices
  - Models probabilities as a mixture of distributions
- Training:
  - Likelihood maximization with EM-algorithm

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{wd} \log p(w|d) \to \max_{\Phi, \Theta}$$

$$\sum_{w} \phi_{wt} = 1; \quad \sum_{t} \theta_{td} = 1$$

Word Embeddings

- Inspired by neural networks for language modeling
- Utilize word-word co-occurrence data
- Skip-Gram model:

$$p(u|v) = \frac{\exp \sum_t \phi_{ut}\theta_{tv}}{\sum_{w \in W} \exp \sum_t \phi_{wt}\theta_{tv}}$$

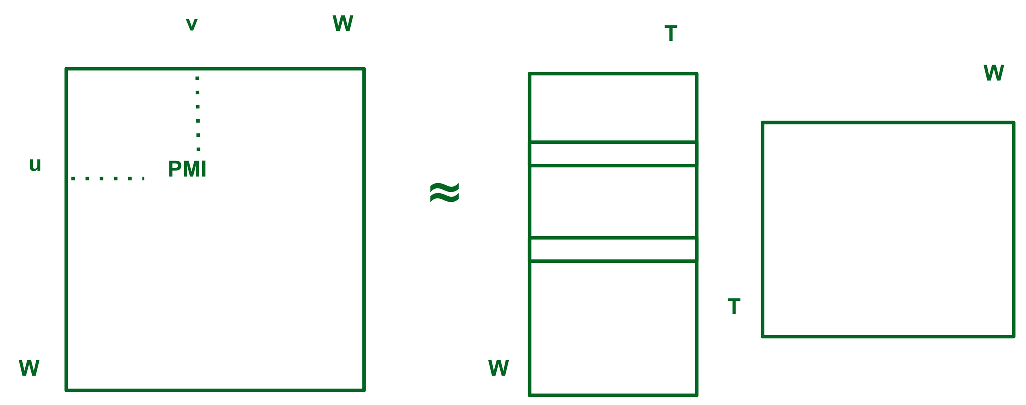
  - Parametrized by two matrices
  - Models probabilities by softmax
- Training:
  - Likelihood maximization with SGD

$$\mathcal{L} = \sum_{v \in W} \sum_{u \in W} n_{uv} \ln p(u|v) \to \max_{\Phi, \Theta}$$

  - No constraints for the parameters

Our approach

- Point out a striking similarity of various models: SGNS, GloVe, PMI-SVD, LDA, NNSE, OIWE....



- Build a hybrid model:

$$p(u|v) = \sum_{t \in T} p(u|t)p(t|v) = \sum_{t \in T} \phi_{ut}\theta_{tv}$$

  - Utilize word-word co-occurrences
  - Parameterize by probabilistic vectors
  - Train with EM-algorithm
- Take advantage of additive regularization of topic modeling to customize the embeddings

ARTM theory

- Additive regularization of topic models:

$$\mathcal{L} + R \to \max_{\Phi, \Theta}; \quad R = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$$

  - Easy way to impose additional requirements
  - Deals with non-uniqueness of matrix factorization
- Multiple modalities (e.g. timestamps, authors, etc):

$$\sum_{m \in M} \lambda_m \underbrace{\sum_{v \in W^0} \sum_{u \in W^m} n_{uv} \ln p(u|v)}_{\text{modality log-likelihood } \mathcal{L}_m(\Phi, \Theta)} \to \max_{\Phi, \Theta},$$

  - Embeds all modalities to the same space
- Examples of regularizers:
  - **Sparsity**: KL-divergence between the topic distributions and uniform distributions
  - **Diversity**: pairwise correlations of the topics
- Implementation: open-source library bigartm.org

Benefits of the two worlds

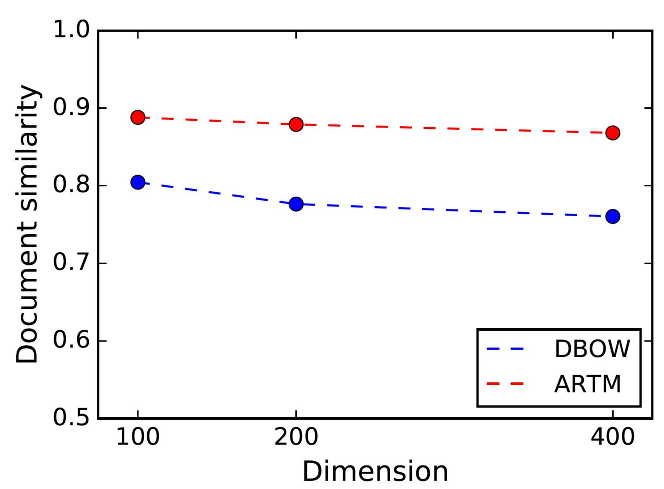
- **Word Similarity - performs on par with SGNS** (on Wikipedia)

Method	WordSim Similarity	WordSim Relatedness	WordSim Joint	Bruni et. al MEN	Radinsky M. Turk
SGNS, *cos*	**0.752**	0.632	0.666	**0.745**	**0.661**
LDA, *hel*	0.530	0.455	0.474	0.583	0.483
ARTM, *dot*	0.728	**0.671**	**0.682**	0.675	0.635

- **Word Analogy - examples:**

Edward + girl - boy	Katherine, Georgina, Susannah, Louise, ...
Alexander + girl - boy	Tamara, Anna, Tatiana, Natasha, Nadia,, ...

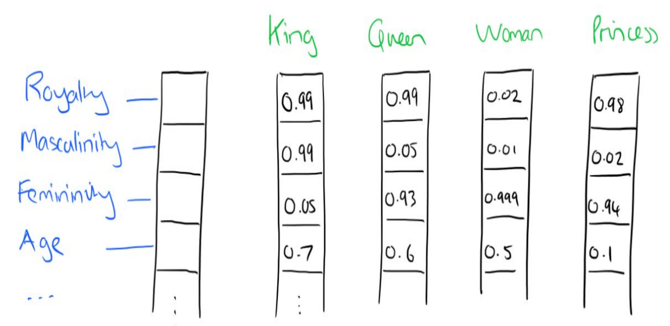
- **Document Similarity - outperforms Paragraph2Vec** (on ArXiv)



Accuracy of choosing the similar paper from a triplet of:
- query paper
- similar paper (by keywords)
- dissimilar paper

Testeset released by Dai et al.

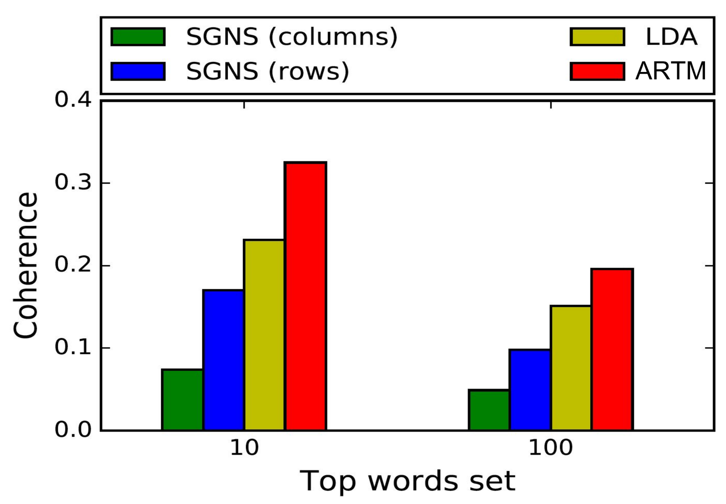
- **Interpretability - drastic improvement**
  - Can components have meaning?



  - How do we measure that?

$$C = \frac{2}{k(k-1)} \sum_{j=2}^{k} \sum_{i=1}^{j} \mathrm{PMI}(w_i, w_j)$$

  - How do the models perform?



Benefits of ARTM

- **Sparsity - 94% of zeros with the KL-regularizer**
- **Multimodal embeddings** (on Lenta.ru news)
  - Meaningful inter-modality similarities - see top-similar words to timestamp embeddings:

2015-12-18 Star Wars Release	2016-02-29 The Oscars	2015-05-09 Victory Day
jedi	statuette	great
sith	award	anniversary
fett	nomination	normandy
anakin	linklater	parade
chewbacca	oscar	demonstration
film series	birdman	vladimir
hamill	win	celebration
prequel	criticism	concentration
awaken	director	auschwitz
boyega	lubezki	photograph

  - Further improvement on word similarity task:

Model	WordSim Similarity	WordSim Relatedness	WordSim+ RG+MC
SGNS	0.630	0.530	0.567
ARTM	0.649	0.565	0.604
Multi-ARTM	**0.682**	**0.580**	**0.611**

Testset translations to Russian are taken from http://russe.nlpub.ru/downloads/

- **Further improvement of interpretability:**

SGNS		ARTM	
transports	rana	art	arbitration
recon	walnut	painting	van
grumman	rashid	museum	requests
convoys	malek	painters	arbitrators
piloted	aziz	gallery	noticeboard

Future work

Apply the proposed probabilistic embeddings to a suite of NLP tasks and take even more advantage of the additive regularization to incorporate task-specific requirements into the model.

References

Anna Potapenko, Artem Popov, Konstantin Vorontsov: Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. To appear in AINL 2017.